EVALUATING THE SENSORY GAP FOR EARTH OBSERVATION IMAGES USING HUMAN PERCEPTION AND AN LDA-BASED COMPUTATIONAL MODEL

*Reza Bahmanyar** *and Ambar Murillo Montes de Oca*

Institute of Remote Sensing Technology (IMF), German Aerospace Center (DLR), Wessling, Germany

ABSTRACT

High resolution Earth Observation (EO) images contain detailed information, making it possible to recognize objects. However, issues such as the sensory gap (the difference between a real life scene and its sensory interpretation) cause difficulties for object recognition. In EO, this gap is rather wide due to sensor resolution, image perspective, scale and field of view (FOV). In this work, human perceptual and computational evaluations of the sensory gap are presented. For the human perceptual evaluation, user labels describing image patch content are gathered and analyzed. Results highlight issues caused by the sensory gap, e.g., FOV (image patch size) limits the contextual clues which can be used to disambiguate objects. The effect of FOV is then computationally analyzed as the difference between the scene context discovered by Latent Dirichlet Allocation from content within a certain FOV and the ground truth. Results indicate that increasing the FOV decreases the sensory gap.

Index Terms— Sensory Gap, Latent Dirichlet Allocation, Human Perception, Earth Observation

1. INTRODUCTION

In the literature, the sensory gap has been defined as the gap between a real life scene, and the information of this scene captured by sensors [1]. A desired object can either be perceived directly by the user, or detected after processing the information in a machine learning application. In either case, a sensory gap exists. Causes behind the sensory gap can lie in the scene (e.g., clutter, occlusion) or on sensor levels (e.g., perspective, resolution, field of view, perceptual spectra). In EO, the sensory gap is rather wide due to sensors (e.g., radar, multi- and hyper-spectral instruments) which record visual information very differently from the human visual system [2].

The sensory gap is affected by the complexities of the EO images, such as the resolution, perspective, or scale of the visual information [3]. The perspective of the images is a particular challenge in EO, since they present a bird's eye view. As described in the "Recognition By Components" theory [4], objects can be segmented into their geometric components ("geons"), and we recognize them based on the identification of their geons and their structural relationships, which we then match to mental representations. Object recognition should be perspective invariant, so long as the structural relationship between geons can be identified from the different perspective. This is not the case when objects are viewed from above, since major object components can be occluded, making it harder to match the object to the stored mental description. Therefore, from this perspective, object identification is more difficult [4].

The sensory gap is also affected by the field of view (FOV) presented. The larger the FOV, the more information is present in the image, specifically contextual information, which the user can apply when trying to identify and recognize an object within the image. Research on object detection and recognition in humans has shown the importance of context [5]. Context can provide information on spatial relations, semantic associations, global scene properties, and pose [6], [7]. When the object is not easily discernible on its own (due to low resolution, for example), contextual information becomes increasingly important [8], [9].

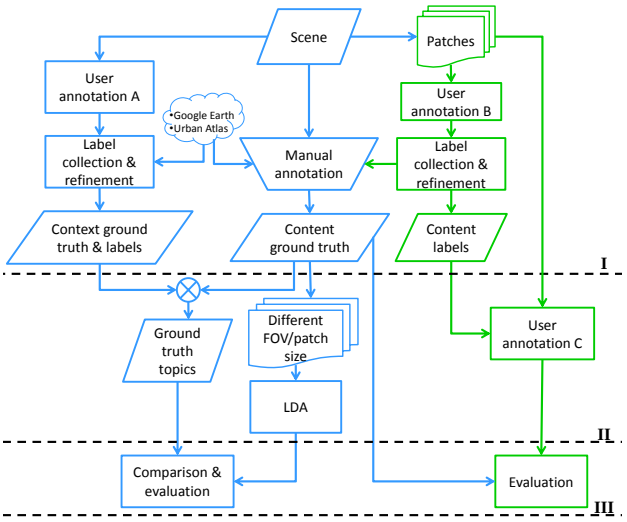


Fig. 1. Process chain of the user perceptual evaluation (shown in green) and the computational evaluation (shown in blue) of the sensory gap.

In this paper, we assess the causes of the sensory gap in EO images by a human perceptual evaluation and a computational evaluation. The human perceptual evaluation is assessed by user labeling (*User annotation C* in Fig. 1) of EO scene image patches, using content labels (listed in Table 2) defined for the scene by a previous annotation (*User annotation B* in Fig. 1). The assigned labels and user feedback are then analyzed (the procedure is shown in green in Fig. 1). Results point to image properties that limit image understanding, such as resolution, which users report is not high enough to readily discriminate objects. Image perspective also presents a challenge, since users are not used to this bird's eye view. The scale of objects in the image patches is also difficult to assess. When users are uncertain of an object's identity, due to other image properties, such as resolution or perspective, they could turn to the context surrounding the object to gather clues to identify it. However, due to the FOV which is constrained by the patch size, users have limited contextual information.

*The author is also affiliated with Munich Aerospace Faculty, Munich, Germany.

Fig. 2. Context ground truth, annotated for the 8 labels in Table 1.

The effect of FOV is then evaluated by a computational method, in which the acquired context from the content ground truth (derived from a manual labeling of the scene described in Section 2) of a certain FOV is statistically analyzed using Latent Dirichlet Allocation (LDA). The sensory gap is considered as the difference between the context of the scene discovered by LDA, based on the occurrence of content ground truth labels within a certain FOV (which corresponds to increasing the image patch size), and the scene ground truth context. The results indicate that increasing the FOV decreases the sensory gap.

The rest of this paper is organized as follows: Section 2 describes the context and content ground truth and label generation process (Phase I in Fig. 1). Section 3 presents the experimental procedures as well as the results and a discussion of the user perceptual and the computational evaluations of the sensory gap (Phases II and III in Fig. 1). Section 4 concludes this paper.

2. CONTEXT AND CONTENT GROUND TRUTH AND LABEL GENERATION

A multi-spectral scene of the Feldmoching area to the north of Munich, Germany, acquired on July 12th, 2010 (10:30 am UT) by the WorldView-2 satellite was used for annotation. The image has a resolution of 1.84 m, was trimmed to a size of 2000×1800 pixels, and three bands were displayed (RGB).

The process chain of the present study (please refer to Fig. 1) shows an overview of the necessary steps starting from the EO *Scene*, using both human user experiments (the process steps shown in green) and computer experiments (shown in blue), to evaluate the sensory gap from both perspectives. In the initial phase, the scene was given to 9 human users (none of whom had a background in image processing), who were asked to annotate the image using the LabelMe tool [10]. This refers to *User annotation A* in the process chain in Fig. 1. In this step, users were presented with the scene, and given a short demo of the tool. A free text annotation [11] was conducted - meaning that users were asked to label what they see, without using references or dictionaries. This approach was selected to gather labels based on user perceptions, without external influences. Each user generated an average of 19 unique labels. In the following step, *Label collection & refinement*, all unique labels (excluding duplicates, plurals, synonyms) were identified, and polygons from the 9 annotations were compared to identify their commonalities. These annotations produced labels corresponding to higher level semantics, such as "industrial areas" and "urban areas", indicating that users focused on the broader "gist" of the scene, as opposed to its details. These higher level semantics were gathered, and loosely refined based on Urban Atlas[1], 8 context labels were determined (please refer to Table 1). These context labels were used together with Google Earth[2] to manually annotate the image and create a *Context ground truth and labels* (please refer to Fig. 2 for a screenshot of this annotation).

1	Agricultural & semi-natural areas	5	Residential areas
2	Industrial/Commercial/Public/Military	6	Sport and leisure
3	Isolated structures	7	Transportation infrastructure
4	Natural areas	8	Water body

Table 1. Context labels

1	Agricultural field	7	Greenhouse	13	Railway
2	Building	8	Highway	14	Road
3	Crop	9	House	15	Soccer field
4	Factory	10	Isolated trees	16	Solar panels
5	Forest	11	Lake	17	Street
6	Grass	12	Parking lot	18	Tennis court

Table 2. Content labels

For the user experiments, the scene was divided into 200×200 patches, with 50% overlap, resulting in 323 *Patches*. Then *User annotation B* was carried out, where 3 different users did a free text annotation [11] labeling an average of 108 patches each. Next, *Label collection & refinement* took place, so that 18 labels describing the content of the patches were left (please refer to Table 2). In contrast to the labels given by the first 9 annotators, these labels corresponded to lower level semantic categories, such as "garden" and "houses". These 18 *Content labels* were used as a lower level semantic dictionary, which provided a manageable set of terms, but was also rich enough to highlight the previously mentioned problems associated with the sensory gap. For example, the labels "road", "street", and "highway" illustrate perception problems due to scale; "agricultural field" and "grass" highlight issues with resolution; "building" and "house" highlight issues with perspective.

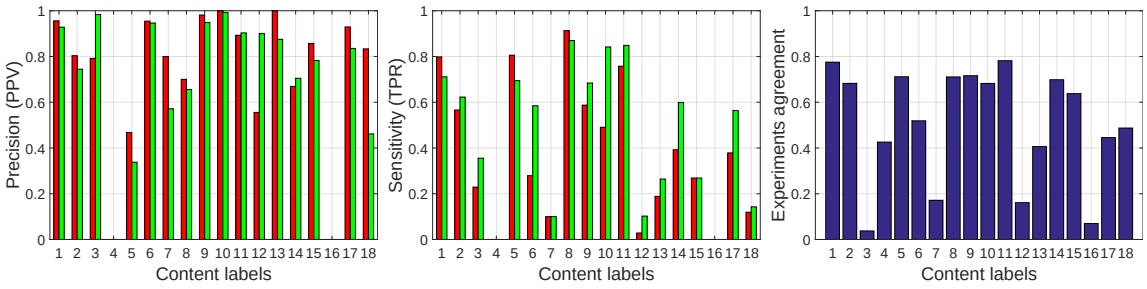
The 18 *Content labels*, together with Google Earth, were used in a manual annotation of the scene, creating a *Content ground truth*. At this point we have both *Context* and *Content ground truth and labels* (please refer to Phase I in Fig. 1). Phase II describes our experimental procedure, which will be detailed in Section 3.1.1 from the user experiment side, and in Section 3.2.2 from the computer side. Phase III consists of the experimental outputs addressing the user perceptual and the computational evaluations of the sensory gap, which will be discussed in Sections 3.1.2 and 3.2.3.

3. THE SENSORY GAP: A USER PERCEPTUAL AND A COMPUTATIONAL EVALUATION

3.1. User perceptual evaluation

For a user perceptual evaluation of the sensory gap, *User annotation C* (please refer to Fig. 1) was carried out. The experimental procedure and results will be discussed below.

[1] http://www.eea.europa.eu/data-and-maps/data/urban-atlas
[2] https://www.google.com/earth/



(a) (b) (c)

Fig. 3. Precision, sensitivity, and agreement of the labels for the two user experiments. In (a) and (b), UX1 and UX2 are depicted by red and green bars, respectively.

3.1.1. Experimental procedure

The 323 image patches previously described were divided into eight groups (seven groups of 40 patches, one group of 43 patches). Users were each given one group of patches, and a handout with the dictionary of content labels listed in Table 2, each assigned to a number code (e.g., 1=Agricultural field). Users were asked to look at each patch (zooming in as needed), and assign it label codes to represent its semantic content. Sixteen users participated (the first 8 corresponding to user experiment 1 (UX1), the second 8 to user experiment 2 (UX2)), so that each group of patches was labeled twice. After labeling, participants were asked to fill out a short questionnaire, to gauge their perceptions on how confident they were of the correctness of their labels, and to give general feedback.

3.1.2. Results and discussion

The similarity between the user and ground truth labels is computed by two measures: *precision* and *sensitivity*, which are formulated for three types of quantities, namely True Positive (TP), False Positive (FP), and False Negative (FN) [12]. While *precision* ($PPV = \frac{TP}{TP+FP}$) indicates the correctness of the user assigned labels to the patches, *sensitivity* ($TPR = \frac{TP}{TP+FN}$) shows the percentage of the ground truth labels which have been identified by the users. In *precision*, for each label, TP and FP are the number of times the label is correctly and incorrectly assigned by the users, respectively. In *sensitivity*, TP and FN are the number of times a ground truth label is identified or missed, respectively.

The average *precision* rates (UX1=73% and UX2=70%) and *sensitivity* rates (UX1=38% and UX2=46%) were similar for both user experiments (please refer to Fig. 3.a,b). The relatively high *precision* rates indicate that when users assigned labels, they were mostly correct. However, there was a large portion of missing labels, as reflected in the *sensitivity* rates. Users reported their confidence in their labeling as an average of 3.7 on a Likert scale (where 1 is not at all confident and 5 is very confident), indicating that they were aware of the potential inaccuracies of their annotations, including the unlabeled objects they could not identify or detect.

When asked to describe the difficulties of the labeling task, users cited problems with understanding the object scales, which then led to questions on how to distinguish semantically related terms (such as "road" and "street") which are typically differentiated by their size. Users also mentioned that the resolution of the image was not high enough to distinguish certain objects. The fact that they could not see the contextual information surrounding the patch, combined with the perspective of the image, made users unsure of what certain objects would look like. Therefore, they would have liked to use examples of labeled patches as a guide.

To further understand the patterns of errors, missing and correctly identified labels, we looked at the *precision* and *sensitivity* of each label, as well as the user feedback given. The results were analyzed with regard to the sensory gap.

In terms of incorrect identification, two object classes stand out: "factory" and "solar panels". Although neither of these object classes were present in the image provided, users detected them. Due to the image's perspective, which does not provide height or depth information, and the human eye is not accustomed to this perspective, the user can only see a cluster of similar buildings. Paths and small parking lots may be confused with factory infrastructure such as pipes connecting different sections of the factory. In the case of solar panels, the effect of perspective resulted in confusing greenhouses with solar panels (probably due to the way they reflect light).

Issues of scale are highlighted with the labels "highway", "road" and "street". "Highway" is more likely to be confused with "road" or "street", than to go undetected. In the case of "street" or "road", users are more likely to miss them or to confuse them with each other. User feedback indicates that these objects are similar and distinguished based on size; however, the limited FOV of the patches makes this difficult for the users to judge.

The average user label agreement rate was found to be 50.6% among all categories (please refer to Fig. 3.c). High agreement on several categories indicates they were easier to detect and discriminate (e.g., "lake" and "agricultural field"). "Solar panels" have a particularly low agreement rate, because this object category was not present in the image, and users confused it with different objects. In the case of "factory", although this object category was not present in the image, the user agreement rate is not as low because users confused the same objects with factory, indicating they have a similar mental representation of what it should look like. Two other categories have a particularly low agreement rate: "greenhouse" and "parking lot". These categories are hard to discriminate or hard to detect, and users mostly missed them, as can be seen by their corresponding *sensitivity* rates. The category "crop" also had a very low agreement rate. Even though most of the labels for "crop" were correctly assigned, users did not label a large percentage of the crops in the image, as evidenced by the *sensitivity* rates. User feedback reported confusion between the categories "crop" and "agricultural field", since the resolution of the image was not high enough for

them to make this distinction. Users also expressed difficulties distinguishing between "building" and "house"; however, there was a high agreement rate for these categories, indicating that even if users express a degree of confusion between the terms, they share a similar mental representation of them.

These results highlight the ways in which the image perspective, resolution, scale and FOV are some of the causes behind the sensory gap from a human user perspective.

3.2. Computational evaluation

In Section 3.1, the sensory gap's causes (image resolution, perspective, scale, and FOV) are explored via a user perceptual evaluation. Among them, in the following sections, FOV is further assessed via a computational evaluation, in which LDA performs a statistical analysis of the contextual clues a given patch with a certain FOV provides (a general overview is shown in Fig. 1).

3.2.1. Latent Dirichlet Allocation

Latent Dirichlet Allocation (LDA) is a statistical generative model, which has been introduced for discovering hidden structures behind collections of text documents, and represents them as mixtures of so-called *topics* [13]. Assuming each document d as a combination of N_d words, $d = \{w_1, w_2, ..., w_{N_d}\}$, LDA discovers the latency as a set of topics, $Z = \{z_1, z_2, ..., z_K\}$, where topics are distributions over a fixed dictionary of words. In a learning phase, LDA finds the posterior distribution, the topic distributions in the documents. Since computing the posterior is intractable, LDA uses approximation inference algorithms such as variational Expectation Maximization.

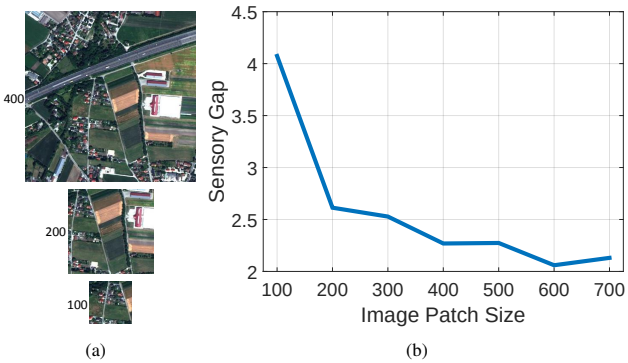
3.2.2. Methodology

As a first step in our experiments, the *Context labels* (please refer to Table 1) are represented as distributions over the *Content labels* (listed in Table 2), $W := \{w_1, w_2, ..., w_n\}$; this representation is called *Ground truth topics*, $\tilde{Z} := \{\tilde{z}_1, \tilde{z}_2, ..., \tilde{z}_m\}$. The distributions are obtained by overlapping the context and the content ground truths, and by pixel-wise measuring the overlap for each pair of context and content labels.

As a next step, the content ground truth is split into patches, where the patch size reflects the patch's FOV. The coverage of the content labels in each patch is considered as the occurrence probability, and this value is used to represent the patch as a histogram of the content labels. LDA is then applied to the histograms to discover the latent topics, $Z := \{z_1, z_2, ..., z_k\}$, behind the patch collection (reflecting the scene context), where each topic is a distribution over the content labels. Since the FOV limits the contextual clues, the resulting scene context differs from the ground truth context which is derived from the complete scene. This difference is then considered as the effect of FOV on the sensory gap. Fig. 4.a exemplifies how changing the FOV limits contextual clues. For a 100 pixel patch, for example, roads cannot be well identified using the contextual clues.

The difference between the two sets of topics is measured by symmetrized Kullback-Leibler divergence [14]:

$$D_{KL}(R_i||Q_j) = \frac{1}{2}[\sum_{x=1}^{n} R_i(x)\ ln\frac{R_i(x)}{Q_j(x)} + \sum_{x=1}^{n} Q_j(x)\ ln\frac{Q_j(x)}{R_i(x)}], \quad (1)$$



(a) (b)

Fig. 4. (a) Example of limitation of the contextual clues by changing the FOV. (b) Influence of FOV (image patch size) on the sensory gap.

where $R_i(x) = p(w_x|\tilde{z}_i)$ and $Q_j(x) = p(w_x|z_j)$. For each LDA-topic, the closest ground truth topic is considered as its corresponding topic. The sum of the distances of the LDA-topics to their corresponding ground truth topics is then computed as the final distance between the two sets of topics, corresponding to the sensory gap.

3.2.3. Results and discussion

In our experiments, LDA is applied to the content label representation of the image patches for various numbers of topics, $k \in [5, 12]$. Since LDA does not provide unique results, each experiment is repeated five times. The final sensory gap for a particular patch size is then obtained by averaging over all of the experiments. As Fig. 4.b shows, increasing the FOV (patch size) significantly reduces the sensory gap up to a certain point (200 pixels). Further increasing the FOV causes no significant change to the sensory gap. This demonstrates that for the given scene considering all its properties (e.g., size, resolution, spectrum) a patch size of more than 200 pixels, statistically, does not add many contextual clues to each patch.

4. CONCLUSION AND FUTURE WORK

In EO, the sensory gap is rather wide due to sensor resolution, image perspective, scale and FOV. In this work, the sensory gap is assessed by human perceptual and computational evaluations. For a human perceptual evaluation, user labels describing image patch content are gathered and analyzed. The results highlight issues caused by the sensory gap. For example, the bird's eye view perspective of the image is one which humans are not accustomed to, and therefore affects object recognition. Resolution and scale present additional difficulties for object recognition. Users can disambiguate objects by gathering context from the image's FOV; therefore, a limited FOV makes issues such as resolution more serious. The effect of FOV on the sensory gap is also assessed via a computational evaluation. There, the sensory gap is defined as the difference between the scene context discovered by LDA from content within a certain FOV (image patch size) and the ground truth context. The results indicate that increasing the FOV decreases the sensory gap. Future work could extend the research on FOV and how it interacts with other factors that cause the sensory gap (such as resolution).

5. REFERENCES

[1] A. W. M. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain, "Content-based image retrieval at the end of the early years," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 12, pp. 1349–1380, 2000.

[2] R. Bahmanyar and M. Datcu, "Measuring the semantic gap based on a communication channel model," in *Proc. 20th IEEE International Conference on Image Processing (ICIP)*, Melbourne, 2013, pp. 4377–4381.

[3] A. Murillo Montes de Oca, N. Nistor, and M. Datcu, "Creating a Reference Data Set for Satellite Image Content Based Retrieval," in *Proc. Conference on Big Data from Space (BiDS)*, Frascati, 2014, pp. 71–75.

[4] I. Biederman, "Recognition-by-components: a theory of human image understanding," *Psychological Review*, vol. 94, no. 2, pp. 115–147, 1987.

[5] I. Biederman, R. J. Mezzanotte, and J. C. Rabinowitz, "Scene Perception: Detecting and Judging Objects Undergoing Relational Violations," *Cognitive Psychology*, vol. 14, no. 2, pp. 143–177, 1982.

[6] C. Green and J. E. Hummel, "Familiar interacting object pairs are perceptually grouped," *Journal of Experimental Psychology. Human Perception and Performance*, vol. 32, no. 5, pp. 1107–19, Oct. 2006.

[7] A. Oliva and A. Torralba, "The role of context in object recognition," *Trends in Cognitive Sciences*, vol. 11, no. 12, pp. 520–7, Dec. 2007.

[8] A. Torralba, "How many pixels make an image?," *Visual Neuroscience*, vol. 26, no. 1, pp. 123–31, 2009.

[9] E. Barenholtz, "Quantifying the role of context in visual object recognition," *Visual Cognition*, vol. 22, no. 1, pp. 30–56, Dec. 2013.

[10] B. C. Russell, A. Torralba, K. P. Murphy, and W. T. Freeman, "LabelMe: A Database and Web-Based Tool for Image Annotation," *International Journal of Computer Vision*, vol. 77, no. 1-3, pp. 157–173, Oct. 2008.

[11] A. Hanbury, "A Survey of Methods for Image Annotation," *Journal of Visual Languages & Computing*, vol. 19, no. 5, pp. 617 – 627, 2008.

[12] Tom Fawcett, "An introduction to ROC analysis," *Pattern Recognition Letters*, vol. 27, no. 8, pp. 861–874, June 2006.

[13] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet Allocation," *The Journal of Machine Learning Research*, vol. 3, pp. 993–1022, Mar. 2003.

[14] S. Kullback and R. A. Leibler, "On information and sufficiency," *The Annals of Mathematical Statistics*, vol. 22, no. 1, pp. 79–86, Mar. 1951.